DOCUMENT RESUME

ED 138 620 TM 006 241

AUTHOR Gillmore, Gerald M.; And Others
TITLE The Teacher and the Course as Units of Analysis in the Generalizability of Student Ratings of Instruction. 77-9.
INSTITUTION Washington Univ., Seattle. Educational Assessment Center.
PUB DATE Feb 77
NOTE 29p.

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Analysis of Variance; *Course Evaluation; Effective Teaching; *Higher Education; Rating Scales; *Reliability; *Student Evaluation of Teacher Performance; *Teacher Evaluation
IDENTIFIERS Generalizability Theory

ABSTRACT

Most previous studies of the dependability of student ratings of instruction have used the class as the unit of analysis, thus confounding the effects of the course, the teacher, and the interaction between the two. This paper presents two studies in which generalizability theory is applied to data collected within designs allowing for independent estimates of the variance components for the teacher effect and for the course effect. Results show that generalizing over items and students but not over teachers or courses yields highly dependable results. Generalizing over courses as well as items and students, with the teacher as the unit of analysis, yields moderately dependable results. Generalizing over teachers, items and students with the course as the unit of analysis yields very low dependability. Implications of the results for tenure and curricular decisions are discussed. (Author)

**********
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
**********

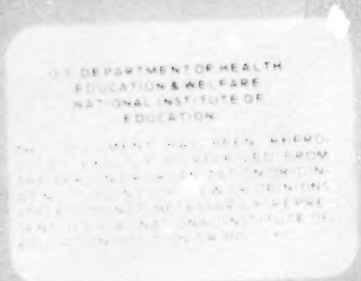
ED138620

# EAC REPORTS

EDUCATIONAL ASSESSMENT CENTER

University of Washington, 1400 Campus Parkway, PB–30, Seattle, Washington 98195 (206) 543–1170

TM006 241




Instructional Evaluation

Educational, Psychological, and Survey Research

Test Administration, Construction, Scoring, and Analysis

2

Educational Assessment Center

University of Washington

February 1977

77-9

# The Teacher and the Course as Units of Analysis in the Generalizability of Student Ratings of Instruction

Gerald M. Gillmore, Michael T. Kane*,

and Richard W. Naccarato

## Abstract

Most previous studies of the dependability of student ratings of instruction have used the class as the unit of analysis, thus confounding the effects of the course, the teacher, and the interaction between the two. This paper presents two studies in which generalizability theory is applied to data collected within designs allowing for independent estimates of the variance components for the teacher effect and for the course effect. Results show that generalizing over items and students but not over teachers or courses yields highly dependable results. Generalizing over courses as well as items and students, with the teacher as the unit of analysis, yields moderately dependable results. Generalizing over teachers, items and students with the course as the unit of analysis yields very low dependability. Implications of the results for tenure and curricular decisions are discussed.

*Michael T. Kane is at the National League for Nursing, 10 Columbus Circle, New York, New York 11375

---

Educational Assessment Center Project: 506

# The Teacher and the Course as Units of Analysis in the Generalizability of Student Ratings of Instruction

Gerald M. Gillmore, Michael T. Kane, and Richard W. Naccarato

Previous studies (Kane, Gillmore, and Crooks, 1976; Gillmore, Kane, and Naccarato, 1976; Kane, Crooks, and Gillmore, 1976) have shown that generalizability theory (Cronbach, Glaser, Nanda, and Rajaratnam, 1972) can be a valuable tool in analyzing the dependability of student ratings of instruction. Specifically, these studies have discussed three distinct generalizability coefficients within the common paradigm under which these ratings are collected. The coefficient which is appropriate in a given setting depends upon whether one wishes to generalize over both students and items, or only one or the other. Furthermore, the characteristics of each of these coefficients can be easily traced as a function of changes in the number of raters and/or items. In addition, the multifaceted interpretation which underlies generalizability theory was useful in showing that the variability of student responses is consistently greater than the variability within students, indicating that increasing the number of students will generally have a much greater impact on dependability than increasing the number of items.

Perhaps the most important practical conclusion arising from the series of analyses reported was that, even for the most conservative (and most often appropriate) coefficient, that is generalizing over both students and items, generalizability coefficients of .70 and greater were found when a class was rated by fifteen or more students, responding to four or more items. This result was reported in the above studies for three distinct instruments in two locations for various disciplinary groupings.

The unit of analysis for the above studies was the class, defined as a particular course (e.g. Psychology 101) taught by a particular teacher. However, the class as a unit of analysis confounds the teacher effect with the course effect. This confounding is acceptable for some uses of student ratings, e.g. feedback to the teacher on his performance in a particular course, but does not permit estimation of the dependability of student ratings for two important applications of such ratings. First in evaluating a teacher's instructional effectiveness for a promotional decision, one would probably want to generalize over all courses that the teacher might teach rather than focusing on results of one or more specific courses. Second, in evaluating a specific curriculum offering, one would probably want to assess the course's rating by students, generalizing over the universe of all possible teachers who might teach it. Analyses in which the teacher and course effects are confounded do not allow an assessment of the dependability of student ratings data for either of these purposes. Additionally, they do not allow a judgment of how many distinct courses taught by a given teacher need to be rated, or how many distinct teachers need to be rated in a course, before a dependable average emerges.

The most obvious solution to the problem of the confounding of these two effects is one which is seldom available within higher education: namely, a factorial design in which teachers are crossed with courses. This requires that the instruction of a set of teachers be rated for the same set of courses. In the rare situation where this might happen the sample sizes for the teacher and course effects are likely to be very small.

One solution to the need for independent assessment of the teacher and course effects is exemplified by Hogan (1973) and Bausel, Schwartz, and Purohit (1975). In both studies, the investigators correlated the ratings of two sections of the same course with the same teacher, two sections of different courses with the same teacher, and two sections of the same course with different teachers. (Bausel et al had two additional conditions which are not of interest here.) The first condition above yields a measure of interrater reliability for course-teacher combinations of classes (see also Costin, Greenough, and Menges, 1972, and Gillmore, 1973). However, the second condition isolates the teacher effect, and the third condition isolates the course effect. The mean correlations reported for the three conditions by Bausel, et al, were .69, .33, and .17 respectively. Median correlations for the first two conditions reported by Hogan were .70 and .40. In the case of two different instructors teaching the same course, Hogan found that the pattern across subscales was inconsistent, but the median correlation was .19.

In the two generalizability studies reported in this paper, generalizability theory is applied to data collected within a design allowing for independent estimates of the variance components for the teacher effect and the course effect. In the first study, courses are nested within teachers thereby providing an estimate of the variance component for the main effect due to teachers, and an estimate of the variance component for the course main effect confounded with the teacher-course interaction. The dependability of the average rating of a teacher generalizing over courses can be estimated from study one.

In the second study teachers are nested within courses, thereby providing

an estimate of the variance component for the main effect due to courses, and an independent estimate of the variance component for the teacher main effect confounded with the teacher-course interaction. The dependability of the average rating of a course generalizing over teachers can be estimated from study two.

Neither study provides an independent estimate of the variance component due to the teacher-course interaction. However, the interaction variance component can be estimated indirectly by combining the results of the two studies.

## Instrument

Data were collected using the University of Washington Instructional Assessment System (IAS) (Gillmore, 1974). The IAS contains five forms with eleven items common to all forms. Since our samples contained users of more than one form, only the responses to the common items were used. All items contain evaluative statements (e.g., "Amount you learned in the course was") followed by six response categories: Excellent, Very Good, Good, Fair, Poor, and Very Poor, coded 5, 4, 3, 2, 1, and 0 respectively.

## Method

### Selection of the Sample

From all classes at the University of Washington in which the IAS was used during the academic year 1974-75, two samples of forty-two pairs of classes were randomly drawn subject to certain restructions. For each class, at least fourteen completed forms had to be present. In order to insure a balanced design, fourteen students were randomly chosen from

classes with more than this number of completed forms.

For study one, only teachers who had been rated in at least two different courses were eligible for inclusion in the sample. If a teacher was rated in two or more courses with at least fourteen students, two of the courses were randomly selected. If a teacher was rated by more than one section within a course, one of these sections was selected randomly. This selection procedure insured a balanced design with courses nested within teachers and students nested within courses.

Similar procedures were followed in study two to insure a balanced design with teachers nested within courses and students nested within teachers. Finally, the sample was drawn for both studies so as to include fourteen pairs of classes from each of three general areas: humanities, social sciences, and physical sciences.

The Design

In the basic design under which the student ratings were collected, each student responded to the same set of items, but a different set of students rated each class. Thus, in study one, students (s) are nested within courses (c), which are, in turn, nested within teachers (t). Items (i) are crossed with all of these factors. Study two is identical except teachers are nested within courses.

The random effects analyses of variance, for both study one and study two, yields estimates for seven components of variance. Expected mean square formulae from whese variance components can be estimated are presented in Table 1 for study one (see Cronbach, et al, 1972, 198-203). The equivalent summary table for study two can be derived by interchanging teacher (t) and course (c) in Table 1.

Insert Table 1 about here

Generalizability Coefficients. Cronbach et al (1972) make an important distinction between studies carried out in order to estimate the generalizability of observations (G studies) and studies upon which decisions will be made (D studies). It is our purpose here to use the variance component estimates which are derived from our G study to forecast the dependability one could expect in a variety of D studies.

The dependability of the results of a D study is expressed by a generalizability coefficient which is defined as the ratio of universe score variance to expected observed score variance. Both the universe score variance and the expected observed score variance depend upon the design and purpose of the D study. For study one, the expected observed score variance with the teacher as the unit of analysis is given as follows (following the notation of Table 1):

$$\varepsilon\,\sigma^2(X_t) = \sigma^2\,(t) + \frac{1}{n_c'}\sigma^2\,(c{:}t) + \frac{1}{n_c'\,n_s'}\sigma^2\,(s{:}c{:}t) + \frac{1}{n_i'}\sigma^2\,(ti)$$

$$+ \frac{1}{n_c'\,n_i'}\sigma^2\,(ci{:}t) + \frac{1}{n_c'\,n_s'\,n_i'}\sigma^2(si{:}c{:}t)\ .$$

As in Table 1, a colon indicates a nesting of effects. For example, $\sigma^2\,(c{:}t)$ is the variance component for courses nested within teachers.

The observed score variance for study one includes all of the estimated variance components, except the variance due to the item main effect, $\sigma^2\,(i)$. In most applications of student ratings, students within all classes will respond to the same set of items, that is, items are crossed with classes rather than being nested within classes, and therefore the item main effect is a constant.

The constants, $n_c'$, $n_s'$, and $n_i'$ are respectively the number of courses, students within each course, and items that are sampled in order to evaluate the teacher in the D study. The observed score is assumed to be the average rating over courses, students, and items, and the contribution of each effect to the observed score variance is inversely proportional to the number of times the effect is sampled in the D study.

The universe score variance depends upon the intended interpretation of the results of the D study, as well as the design of the D study. The decision maker who uses the results of the D study intends to generalize to some universe of acceptable observations, the universe of generalization. The universe score is the expected value of the observed score, taken over the universe of generalization. The universe of generalization, then, can be defined by deciding whether the conditions of each facet in the D study are to be considered a sample from the universe of possible conditions for the facet, or are to be taken as the universe of possible conditions for the facet.

Thus, for each facet the conditions in the D study can be taken as a sample from some larger set of conditions (random effect) or can be said to define the full set of conditions for the facet (fixed effect). Since there are three facets for each of our studies, eight possibilities exist, with each leading to a different definition of the universe score and to a different generalizability coefficient. However, the design of our G study eliminates three of these possibilities. The coefficient which arises when the investigator chooses not to generalize over any of the facets is not estimable for the data from our G study (Kane and Brennan, in press),

because no independent estimate of error variance exists. The two coefficients in each study which would arise from considering courses (or teachers) as random and students as fixed are not meaningful because students were nested within courses (or teachers) in the G study. Thus, students cannot be fixed if courses (or teachers) are random. Each new course (or teacher) sampled must contain a new sample of students.

Of the five remaining generalizability coefficients, two are of particular interest. Kane, Gillmore, and Crooks (1976) and Gillmore et al (1976) argued that in attempting a general assessment of teaching effectiveness, it is appropriate to generalize over both items and students, and in this discussion, we shall assume generalization over these two facets. In study one, the two coefficients which generalize over both students and items are $\varepsilon\rho^2$ (C,S,I) and $\varepsilon\rho^2$ (C*,S,I). We follow the notation of Kane and Brennan (in press) and designate facets which are not being generalized over with an asterisk. The formulae for these two coefficients are as follows:

$$\varepsilon\rho^2\ (C,S,I) = \frac{\sigma^2\ (t)}{\varepsilon\sigma^2\ (X_t)} \qquad (2)$$

$$\varepsilon\rho^2\ (C^*,S,I) = \frac{\sigma^2\ (t) + \frac{1}{n_c'}\,\sigma^2\ (c{:}t)}{\varepsilon\sigma^2\ (X_t)}$$

$\varepsilon\rho^2$ (C,S,I) is approximately equal to the expected value of the correlation of two evaluations based on independent random samples of $n_c'$ courses, $n_s'$ students per course, and $n_c'$ items. This coefficient would be appropriate if the decision maker wished to generalize to all courses the teacher might teach, all students who might enroll in the courses and all items within the

domain which could be used to rate the teacher. If one did not wish to generalize to the entire set of courses, but only to the particular courses included in the D study, then $\varepsilon\rho^2$ (C*,S,I) would be appropriate. $\varepsilon\rho^2$ (C*,S,I) is approximately equal to the expected value of the correlation of two evaluations based on the same $n_c'$ courses, and random samples of $n_i'$ items and $n_s'$ students per course.

In study two, the unit of analysis is the course. The expected observed score variance for study two is given as follows:

$$\varepsilon\rho^2 (X_c) = \sigma^2(c) + \frac{1}{n_t'}\sigma^2 (t:c) + \frac{1}{n_t' n_s'}\sigma^2 (s:t:c) + \frac{1}{n_i'}\sigma^2 (ci)$$
$$+ \frac{1}{n_t' n_i'}\sigma^2 (ti:c) + \frac{1}{n_t' n_s' n_i'}\sigma^2 (si:c:t) \quad . \qquad (4)$$

If we wish to generalize over teachers, the appropriate generalizability coefficient is:

$$\varepsilon\rho^2 (T,S,I) = \frac{\sigma^2(c)}{\sigma^2 (X_c)} \quad . \qquad (5)$$

This coefficient is approximately equal to the expected value of the correlation between two evaluations based on independent random samples of $n_t'$ teachers, $n_s'$ students per teacher, and $n_c'$ items. If we do not wish to generalize over teachers, the appropriate generalizability coefficient is:

$$\varepsilon\rho^2 (T^*,S,I) = \frac{\sigma^2(c) + \frac{1}{n_t'}\sigma^2(t:c)}{\varepsilon\rho^2 (X_c)} \quad . \qquad (6)$$

In this case, the coefficient is approximately equal to the expected value of the correlation between two evaluations based on the same $n_t'$ teachers, and random samples of $n_i'$ items and $n_s'$ students per teacher.

$\varepsilon\rho^2$ (C*,S,I) and $\varepsilon\rho^2$ (T*,S,I) are equivalent to each other and to the $\varepsilon\rho^2$ (S,I) of Kane, Gillmore, and Crooks (1976) and Gillmore et al (1976) in the special case where $n_c'$ is set equal to one for $\varepsilon\rho^2$ (C*,S,I) and $n_t'$ is set equal to one for $\varepsilon\rho^2$ (T*,S,I). That is, in this special case the unit of analysis reverts back to being the class rather than the course or the teacher.

## Results

Study One: The results of the analysis of variance for study one, courses nested within teachers, are presented in Table 2.

---

Insert Table 2 about here

---

Of special interest are the magnitudes of the variance component estimates. The variance component for students nested within classes and for the item by student interaction (confounded with random error) are both large. The variance components for items, the item by course interaction, and the item by teacher interaction are all close to zero. The results are consistent with those reported by Kane, Gillmore, and Crooks (1976) and Gillmore et al (1976). The estimated variance components for teachers and for courses within teachers are both of intermediate magnitude.

Generalizability coefficients are presented in Table 3 for study one. Using Equations (2) and (3), generalizability coefficients can be estimated for any number of courses, items, and students within courses. We chose to present values corresponding to one, two, five, and ten courses and for two levels of students: twenty students per course to represent average size

sections and five students per course to represent very small sections. The generalizability coefficients which we are considering are not influenced greatly by increases in numbers of students beyond twenty. Coefficients are not influenced greatly by the number of items. Thus, we arbitrarily chose ten items as a representative number and based all our computations on this value. The interested reader can use Equations (2) and (3) and the variance component estimates to compute generalizability coefficient values for any size of $n_s'$, $n_i'$ and $n_c'$. The values of the two generalizability coefficients under the above conditions are presented in Table 3.

---

Insert Table 3 about here

---

Considering the values presented in Table 3 when $n_s = 20$, the results of study one show that for a general measure of teaching effectiveness, if one chooses to generalize over courses then a sample of five to ten average-sized courses yields dependable results. If a teacher happens to be teaching smaller classes, a few more courses are necessary to assure dependability. The number of students who rate the various courses seems to make little difference when the number of courses considered is ten or more. Clearly for any number of students, dependability is inadequate when the ratings of only one or two courses are considered. $\varepsilon\rho^2$ (C,S,I) = .40 and .57 for $n_c = 1$ and 2, respectively, when $n_s'$ and $n_i'$ are set equal to infinity. However, if one chooses not to generalize over courses, ratings from one class have adequate dependability, unless that class is small. If the universe of generalization consists of five or more courses, the number

of students per course does not have to be large to assure dependability.

Study Two: The results of the analysis of variance for study two, teachers nested within courses, are presented in Table 4.

---

Insert Table 4 about here

---

Except for $\sigma^2$ (c) and $\sigma^2$ (t:c), the estimated variance components in Table 4 follow the same pattern as those in Table 2. However, the course variance, $\sigma^2$ (c), which was not estimated in study one is close to zero, while the variance due to the teacher effect, confounded with the course-teacher interaction, is relatively large.

Selected generalizability coefficients based on study two results are presented in Table 5. As in the results of study one, we present values for four levels of teachers (one, two, five, and ten), two levels of students (five and twenty) and one level of items (ten).

---

Insert Table 5 about here

---

The results show that if one wishes to evaluate courses by generalizing over teachers, no dependable result is possible for any reasonable sample size of teachers. If one does not wish to generalize over teachers, adequate dependability is attainable unless the class sizes are quite small. Interestingly, the number of teachers averaged over appears to be a negligible factor in determining the magnitude of $\varepsilon\rho^2$ (T*,S,I). This is not surprising since $\sigma^2$ (t:c) is included in both the universe score variance and the expected observed score variance for this coefficient.

Study one and study two: As mentioned earlier, when we choose not to generalize over courses (study one) or teachers (study two) and sample only one condition for these facets we are effectively evaluating particular course-teacher combinations, or classes in our terminology. Therefore, $\varepsilon\rho^2$ (C*,S,I) and $\varepsilon\rho^2$ (T*,S,I) are equivalent estimates when $n_c' = n_t' = 1$ and $n_s'$ and $n_i'$ are equal within each sample. Our results show the two values to be .87 and .85 for study one and study two respectively, when $N_s = 20$, and .65 and .61 respectively when $n_s' = 5$. These differences seem well within the bounds of sampling fluctuation, and give no reason to suspect the samples of the two studies differ in any systematic way.

By combining the results of the two studies, the total class variation can be partitioned into that attributable to teachers, to courses, and to the teacher by course interaction. From study one, we see that 40% of the estimated class variance component, $\hat{\sigma}$ (t) + $\hat{\sigma}$ (c:t), is attributable to the teacher effect. Similarly, we can estimate from study two results that 6% of the class variation is attributed to the course effect. This leaves approximately 54% for the interaction effect. However, this variability is not entirely due to an interaction between courses and teachers. Some is attributable to what would be a section within classes effect; that is the variability of the average ratings of the same teacher, teaching different sections of the same course. Our data confound these two effects since we effectively have only one entry per cell. However, Kane, Crooks, and Gillmore (1976) presented data for which the estimates of the section variance component were negligible for sections within three different courses using the Illinois Course Evaluation Questionnaire. These results

suggest that the variation attributable to a course by teacher interaction is probably not due to any counfounded section effect.

Comparisons with Bausel et al (1975) and Hogan (1973)

It is not possible to directly compare the above results with those obtained by Bausel et al and Hogan, since the number of students in the classes they sampled varied. Both limited their samples to classes rated by at least five students and Bausel et al indicated that their average class size was 21.4. Hogan does not indicate the average class size for his study--this is unfortunate since dependability estimates are dependent on class size, but also points out an advantage of generalizability theory. Once the G study is completed and the variance components are estimated, dependability estimates can be offered for any sample size of a given facet.

Furthermore, the methodology of Hogan and Bausel et al does not examine the generalizability over items, since all classes were rated with the same set of items. Not generalizing over items when using student ratings as a measure of general teaching effectiveness does not very seriously inflate the estimated dependability of the average ratings (see Kane, Gillmore, and Crooks, 1976 and Gillmore, et al, 1976). However, as stated earlier, it does seem appropriate to generalize over both facets.

Finally, the methodology of Bausel, et al, and Hogan restricts them to sample sizes of one teacher, course, or class. Their analyses do not provide estimates of the dependability of averages based on greater numbers of teachers or courses.

In spite of these limitations, it is still of interest to compare

their results with ours. In doing so, we shall arbitrarily assume an average class size ($n_s' = 20$), and an instrument with ten items ($n_i' = 10$). Each of their conditions is listed in Table 6, along with the comparable generalizability coefficient, and the results obtained in each of the three studies. The generalizability coefficients presented in Table 6 assume that items are fixed. Although we believe that coefficients assuming generalization over items are more appropriate for evaluating the dependability of student ratings, the analyses of Bausel et al and Hogan treated items as fixed, and thus generalizability coefficients that do not generalize over items are more comparable.

Insert Table 6 about here

All three studies agree in showing that just generalizing over students yields highly dependable results, generalizing over courses and students yields moderately dependable results, and generalizing over teachers and students yields non-dependable results. The differences that do appear across studies in Table 6 may be attributable to a number of causes. The three studies were conducted in three quite dissimilar institutions, using three different instruments. The number of items and the number of students per class were not the same for the studies. Given these dissimilarities in addition to expected sampling fluctuations, we feel that the comparable results of the three studies show surprising similarity.

## Discussion

With few exceptions, studies presenting reliability estimates for student ratings of instruction have used classes as their unit of analysis,

thereby confounding the teacher, the course, and the interaction between the two. This is not necessarily inappropriate and does not present a serious problem if one's purpose dictates that neither courses or teachers be generalized over. However, this procedure yields inflated estimates when it is appropriate to generalize over either courses or teachers, especially when one wishes to evaluate a course by generalizing over teachers.

Within higher education, using student ratings for making decisions about teachers is much more common than using such ratings for making decisions about courses. The most important decisions made about teachers are those of promotion and tenure. A promotion or tenure decision may imply a long-term commitment to a faculty member and the courses a faculty member teaches are unlikely to remain static over long periods of time. Therefore, it seems appropriate to generalize over all courses he/she might be teaching in evaluating his/her teaching effectiveness. This argues for $\varepsilon\rho^2$ (C,S,I) as the most appropriate index of dependability. If so, our results indicate that data should be collected from as many courses as possible, and a decision based on the average of less than five courses is probably questionable. More courses are required for dependability if the teacher is rated by students in relatively small sections.

On the other hand, if it is fairly certain that the teacher will teach only the specific courses which have been rated, $\varepsilon\rho^2$ (C*,S,I) may be fully appropriate as an index of dependability. In such a case the achievement of adequate dependability is not a problem unless results are available from only one small class.

Our results indicate that a course cannot be dependably evaluated by

student ratings if the decision-maker wishes to generalize over teachers. $\epsilon\rho^2$ (T,S,I) is small for all reasonable values of $n_c'$. If a course is being evaluated as it is taught by a particular teacher or teachers, and there is no desire to generalize beyond this teacher or these teachers, then $\epsilon\rho^2$ (T*,S,I) is appropriate and adequate dependability is achievable.

The course does not seem to be a major factor in the determination of the ratings of a class. Results from our samples do not indicate that some courses are inherently less favorably rated than others. However, a major source of error in the ratings of teachers when generalizing over courses appears to be the teacher by course interaction. The existence of a large interaction component has two practical implications. First, teachers should not be allowed to choose the courses from which ratings are considered in promotion and tenure decisions. Otherwise, judicious selection can easily lead to an overly optimistic view of teaching effectiveness. Secondly, the existence of an interaction suggests that the quality of teaching as perceived by students can be improved by careful assignment of teachers to courses.

## References

Bausell. R. B., Schwartz, S. and Purohit, A. An examination of the conditions under which various student rating parameters replicate across time. Journal of Educational Measurement, 12, 1975, 273-280.

Costin, F., Greenough, W. T., and Menges, R. J. Student ratings of college teaching: Reliability, validity, and usefulness. Review of Educational Research, 1971, 41, 511-535.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratram, N. The Dependability of Behavioral Measurements, New York: John Wiley and Sons, 1972.

Gillmore, G. M. A brief description of the University of Washington Instructional Assessment System. EAC Reports, Educational Assessment Center, University of Washington, 1974 (mimeo).

Gillmore, G. M. Estimates of reliability coefficients for items and subscales of the Illinois Course Evaluation Questionnaire (Research Report 341) University of Illinois, Measurement and Research Division, 1973.

Gillmore, G. M., Kane, M. T., and Naccarato, R. W. The generalizability of student instructional ratings: general theory and application to the University of Washington Instructional Assessment System. EAC Reports, Educational Assessment Center, University of Washington, 1976 (mimeo).

Hogan, T. P. Similarity of student ratings across instructors, courses, and time. Research in Higher Education, 1, 1973, 149-154.

Kane, M. T. and Brennan, R. L. The generalizability of class means. Review of Educational Research (In press).

Kane, M. T., Crooks, T. J., and Gillmore, G. M. Generalizability and the interpretation of student evaluations of teaching. AERA, 1976.

Kane, M. T., Gillmore, G. M., and Crooks, T. J. Student evaluations of teaching: The generalizability of class means. Journal of Educational Measurement, 1976, 13, 171-184.

Table 1

Analysis of Variance for a Balanced Design with Students (S) Nested within Courses (C) Courses Nested within Teachers ( ), and Items (i) crossed with Students (Study 1)*

| Source | Confounded Factors | Notation | Degrees of Freedom | Expected Mean Squares (All Effects Random) |
|---|---|---|---|---|
| t | | t | $n_t - 1$ | $\sigma^2(si:c:t) + n_i\,\sigma^2(s:c:t) + n_s\,\sigma^2(ci:t) + n_s n_i\,\sigma^2(c:t) + n_c n_s\,\sigma^2(ti) + n_c n_s n_i \sigma^2(t)$. |
| Within t | c, tc | c:t | $n_t(n_c-1)$ | $\sigma^2(si:c:t) + n_i\,\sigma^2(s:c:t) + n_s\,\sigma^2(ci:t) + n_s n_i\,\sigma^2(c:t)$ |
| Within tc | s,ts,cs,tcs | s:c:t | $n_t n_c(n_s-1)$ | $\sigma^2(si:c:t) + n_i\,\sigma^2(s:c:t)$ |
| i | | i | $n_i - 1$ | $\sigma^2(si:c:t) + n_s\,\sigma^2(ci:t) + n_c n_s\,\sigma^2(ti) + n_i n_c n_s\,\sigma^2(t)$ |
| ti | | ti | $(n_t-1)(n_i-1)$ | $\sigma^2(si:c:t) + n_s\,\sigma^2(ci:t) + n_c n_s\,\sigma^2(ti)$ |
| Within ti | ci,tci | ci:t | $n_t(n_c-1)(n_i-1)$ | $\sigma^2(si:c:t) + n_s\,\sigma^2(ci:t)$ |
| Within tci | si,tsi,csi, tcsi | si:c:t | $n_t n_c(n_s-1)(n_i-1)$ | $\sigma^2(si:c:t)$ |

* $n_t$ = number of teachers $n_c$ = number of courses $n_s$ = number of students $n_i$ = number of items

Table 2

Study One Analysis of Variance Summary Table

| Source | SS | DF | MS | Estimated Variance Component |
|---|---|---|---|---|
| t | 2150.75 | 41 | 52.46 | .08 |
| c:t | 1069.78 | 42 | 25.47 | .12 |
| s:c:t | 6451.36 | 1092 | 5.91 | .49 |
| i | 297.24 | 10 | 29.72 | .02 |
| ti | 788.75 | 410 | 1.92 | .03 |
| ci:t | 401.47 | 420 | 0.96 | .03 |
| si:c:t* | 5496.71 | 10,920 | 0.50 | .50 |

*This source of variance includes the contribution of random error

Table 3

Estimated Generalizability Coefficients for Various Conditions Within Study One

| | $\varepsilon\rho^2$ (C,S,I) | | $\varepsilon\rho^2$ (C*,S,I) | |
|---|---|---|---|---|
| $n'_c$ | $n'_s = 5$ | $n'_s = 20$ | $n'_s = 5$ | $n'_s = 20$ |
| 1 | .26 | .35 | .65 | .87 |
| 2 | .40 | .53 | .71 | .93 |
| 5 | .62 | .71 | .81 | .95 |
| 10 | .82 | .84 | .95 | .97 |

$n_i = 10$ for all conditions

Table 4

Study Two Analysis of Variance Summary Table

| Source | SS | DF | MS | Estimated variance Component |
|---|---|---|---|---|
| c | 1352.21 | 41 | 32.98 | .01 |
| t:c | 1287.25 | 42 | 30.65 | .16 |
| s:t:c | 6261.81 | 1092 | 5.73 | .48 |
| i | 158.44 | 10 | 15.84 | .01 |
| ci | 573.31 | 410 | 1.40 | .01 |
| ti:c | 437.68 | 420 | 1.04 | .04 |
| si:t:c* | 5197.48 | 10,920 | 0.47 | .47 |

*This source of variance includes the contribution of random error

Table 5

Estimated Generalizability Coefficients for Various Conditions Within Study Two

| | $\varepsilon\rho^2$ (T,S,I) | | $\varepsilon\rho^2$ (T*,S,I) | |
|---|---|---|---|---|
| $n'_i$ | $n'_s = 5$ | $n'_s = 20$ | $n'_s = 5$ | $n'_s = 20$ |
| 1 | .04 | .05 | .61 | .85 |
| 2 | .07 | .10 | .62 | .85 |
| 5 | .16 | .20 | .65 | .86 |
| 10 | .27 | .33 | .69 | .86 |

$n'_i = 10$ for all conditions

Table 6

Comparison of the Present Study with the Results of Bausel et al (1975) and Hogan (1973)

| Bausel et al and Hogan condition | Comparable Generalizability Coefficient | Average Correlations | | Generalizability Coefficient Value* |
|---|---|---|---|---|
| | | Bausel et al | Hogan | |
| Same Teacher, Same Class | $\varepsilon\rho^2(T^*,S,I^*)$ <br> $\varepsilon\rho^2(C^*,S,I^*)$ | .69 | .70 | .80 or .76 |
| Same Teacher, Different Class | $\varepsilon\rho^2(C,S,I^*)$ | .33 | .40 | .48 |
| Different Teacher, Same Class | $\varepsilon\rho^2(T,S,I^*)$ | .17 | .19 | .05 |

*$n'_c = n'_t = 1$

$n'_s = 20$

$n'_i = 10$